COMPSCI 688: Probabilistic Graphical Models

Lecture 19: Black-Box Stochastic Variational Inference

VI

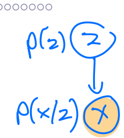
Dan Sheldon

Manning College of Information and Computer Sciences
University of Massachusetts Amherst

Partially based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)

Review

Variational Inference



Have: $p(z,x)$, fixed x, $p(x)$ "hard"

Want: $p(z|x) \approx q_\phi(z)$

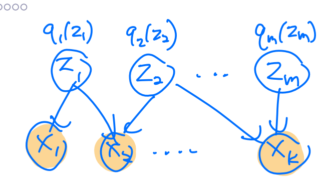
target posterior; approximating/variational dist

1. **Input**: $p(z,x)$ and fixed x
2. Choose some approximating family $q_\phi(z)$
3. Maximize ELBO(ϕ) wrt ϕ — equivalent to minimizing KL($q_\phi(z) \,\|\, p(z|x)$)
4. Use $q_\phi(z)$ as a proxy for $p(z|x)$

$$\text{ELBO}(\phi) = \mathbb{E}_{q_\phi(Z)}\left[\log \frac{p(Z,x)}{q_\phi(Z)}\right] = \mathbb{E}_{q_\phi(Z)}\left[\log p(Z,x)\right] - \mathbb{E}_{q_\phi}\left[\log q_\phi(Z)\right]$$

$$= \sum_z q_\phi(z) \log \frac{p(z,x)}{q_\phi(z)}$$

Variational Inference



causes diseases

effects symptoms

Something we skipped: $p(z,x)$ discrete graphical model, $q(z) = \prod_j q_j(z_j)$ ("mean field"). "coordinate ascent VI", CAVI

Today: z continuous, $p(z,x)$ black box, $q(z)$ TBD

Motivation: Continuous Latent Variable Models

Factor Analysis (PCA)

image
audio clip
feature vector

Factor analysis is a classical statistical model. It posits an observed vector $\mathbf{x} \in \mathbb{R}^d$ is generated as a linear combination of basis vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ with weights $z_1, \ldots, z_m$ plus noise: $x = W\ z + \text{noise}$

$$\underset{d}{\begin{bmatrix} x \end{bmatrix}} = \begin{bmatrix} w_1 & w_2 & \cdots & w_m \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix} + \text{noise}$$

$$x = z_1 \cdot \vec{w}_1 + \ldots + z_m \cdot \vec{w}_m + \text{noise}$$

factor weights

$$x = Wz + \text{noise}$$

"dictionary of basis elements"

$x = Wz + \text{noise}$ $\quad$ $\text{noise} \sim \mathcal{N}(0, \Psi)$

$x \sim \mathcal{N}(Wz, \Psi)$

Probabilistic factor analysis assumes the weights are drawn from a standard normal. The generative process is:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, I)$$
$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mathbf{W}\mathbf{z}, \Psi)$$

noise covariance, often $\sigma^2 I$

(Typically Ψ is diagonal and the data is pre-processed so $\mathbf{x}$ has zero mean.)

$p(z|x)$

Visualization: PCA Demo

Factor Analysis: Learning

$z \rightarrow x$

Consider learning the parameters $\theta = (\underset{\text{dictionary}}{\mathbf{W}}, \underset{\text{noise covariance}}{\Psi})$ given data $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ assumed to be independently drawn from this model.

Since $\mathbf{z}$ is latent, the log-likelihood of a single datum $\mathbf{x}$ is $\log p(\mathbf{x})$, the "log-marginal likelihood".

In this model, the marginal likelihood is available in *closed form*:

$$p(\mathbf{x}) = \int \mathcal{N}(\mathbf{z}; 0, I)\mathcal{N}(\mathbf{x}; \mathbf{Wz}, \Psi)d\mathbf{z} = \boxed{\mathcal{N}(\mathbf{x}; 0, \mathbf{WW}^\top + \Psi)}$$

$\int p(z) \quad p(x|z) \quad dz$

max $\log p(x)$ w.r.t. W, Ψ

Therefore, we can learn by maximizing the log-marginal likelihood:

$$\mathcal{L}(\theta) = -\frac{N}{2}\log(|2\pi\Sigma|) - \frac{1}{2}\sum_{n=1}^{N} \mathbf{x}^{(n)\top}\Sigma^{-1}\mathbf{x}^{(n)}, \qquad \Sigma = \mathbf{WW}^\top + \Psi$$

Alternately, there is an EM algorithm for this model where both E and M steps have simple forms.

Takeaways:
- useful model
- inference/learning "easy"

Factor Analysis: Generalizations

This model is "easy", but factor analysis has many generalizations that make exact learning and inference intractable.

A variational autoencoder (VAE) uses a nonlinear-function f_θ instead of a linear transformation $\mathbf{W}$ to map from $\mathbf{z}$ to the mean of $\mathbf{x}$:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, I)$$
$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; f_\theta(\mathbf{z}), \Psi)$$

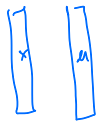
$z \xmapsto{W_2} \mu$

$z \xmapsto{f_\theta(z)} \mu$

A typical structure for f_θ is a multi-layer neural network, e,g.

$$f_\theta(\mathbf{z}) = h_2(\mathbf{b}_2 + \mathbf{W}_2 \cdot \overbrace{h_1(\underbrace{\mathbf{b}_1 + \mathbf{W}_1\mathbf{z}})}^{z'})$$

where h_2, h_1 are element-wise nonlinear functions.



Another generalization changes the likelihood, e.g., to a Bernoulli distribution:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, I)$$
$$p(x_j|\mathbf{z}) = \text{Bernoulli}(x_j; \underbrace{(f_\theta(\mathbf{z}))_j}^{\mu_j}), \qquad j = 1, \ldots, d$$

$$z \xmapsto{W_z} \mu \xmapsto{\mathcal{N}} x$$
$$z \xmapsto{f_\theta(z)} \mu \xmapsto{\text{Bernoulli}} x$$

Inference and Learning in Generalized Models

Almost any change from the basic factor analysis model makes it so we can't compute the marginal likelihood $p(\mathbf{x})$ exactly, so inference and learning become hard.

The model is *only* tractable with linear transformations and a Gaussian likelihood.

We need additional inference tools for the generalizations.

Black-Box Stochastic Variational Inference

Black-Box Stochastic Variational Inference

BBSVI ~ statistical models,
e.g. PPL NumPyro

A general inference approach that works well for models with continuous latent variables, including factor analysis, is *black-box stochastic variational inference*:

- **Black box**: only requires computing $\log p(z, x)$ and its gradients for different z
- **Stochastic**: optimizes the ELBO using Monte Carlo estimates

Stochastic Variational Inference

"SGD" for ML
– subsample data

model
whole dataset

- **Input**: $p(z, x)$ and fixed x, family q_ϕ
- Start with some ϕ
- For $t = 1, 2, \ldots, T$
  - $g \leftarrow$ **unbiased estimate** of ∇_ϕELBO
  - Take a small step: $\phi \leftarrow \phi + \epsilon g$ (or Adam or other optimizer)
- **Return** ϕ, use q_ϕ

Main issue: how to get g?

$g \approx \nabla_\phi \text{ELBO}(\phi)$

unbiased $\mathbb{E}[g] = \nabla_\phi \text{ELBO}(\phi)$

Warmup: Estimating the ELBO

$\mathbb{E}_{q_\phi(z)}\left[\log \frac{p(z,x)}{q_\phi(z)}\right]$

It's easy to estimate the ELBO via Monte Carlo samples:

$$\text{ELBO}(\phi) \approx \frac{1}{K}\sum_{i=1}^{K} \log \frac{p(z^{(i)}, x)}{q_\phi(z^{(i)})} \qquad z^{(i)} \sim q_\phi$$

This is unbiased: expected value of RHS = ELBO(ϕ) for any value of K.

Estimating the Gradient? (False Start)

What happens if we try to estimate the gradient the same way? We want:

$$\nabla_\phi \text{ELBO}(\phi) = \nabla_\phi \mathbb{E}_{q_\phi}\left[\log \frac{p(Z,x)}{q_\phi(Z)}\right]$$

Consider the estimate: $(K=1)$

$$\nabla_\phi \left(\log \frac{p(z,x)}{q_\phi(z)}\right), \quad z \sim q_\phi.$$

This is not unbiased! It neglects the fact that the *distribution* of z depends on ϕ:

$$\nabla_\phi \text{ELBO}(\phi) = \nabla_\phi \int q_\phi(z) \log \frac{p(z,x)}{q_\phi(z)} dz$$

The false start incorrectly interchanges the gradient and the expectation:

$$\nabla_\phi \mathbb{E}_{q_\phi}\left[\log \frac{p(Z,x)}{q_\phi(Z)}\right] \neq \mathbb{E}_{q_\phi}\left[\nabla_\phi \log \frac{p(Z,x)}{q_\phi(Z)}\right]$$

$$= \int \nabla_\phi \left(q_\phi(z) \log \frac{p(z,x)}{q_\phi(z)}\right) dz$$

"score function estimator"

Reparameterization Trick

$\mathbb{E}_{q_\phi}[\cdots] \xrightarrow{\text{trick}} \mathbb{E}_{q(\epsilon)}[\quad]$

The reparameterization trick is a way to convert the ELBO into an expectation with respect to a *fixed* distribution (independent of ϕ) so we can interchange the gradient and expectation. The idea is to draw samples of z by transforming a random variable from a fixed base distribution.

Example: $z = \mu + \sigma\epsilon,\ \underbrace{\epsilon \sim \mathcal{N}(0,1)}_{\text{fixed base}} \implies z \sim \mathcal{N}(\mu, \sigma^2)$

General case: $z = \mathcal{T}_\phi(\epsilon),\ \epsilon \sim q(\epsilon) \implies z \sim q_\phi(z)$

We call $\mathcal{T}_\phi$ and $q(\epsilon)$ a *reparameterization* of q_ϕ

Generate $\epsilon \sim q(\epsilon)$, compute $z = \mathcal{T}_\phi(\epsilon)$

ELBO Gradient with Reparameterization

$\mathbb{E}_{q_\phi(z)}\left[\log \frac{p(z,x)}{q_\phi(z)}\right]$

With reparameterization, we can write the ELBO as an expectation over $q(\epsilon)$:

$$\text{ELBO}(\phi) = \mathbb{E}_{q(\epsilon)}\left[\log \frac{p(\overset{z}{\mathcal{T}_\phi(\epsilon)}, x)}{q_\phi(\underset{z}{\mathcal{T}_\phi(\epsilon)})}\right]$$

Now we can interchange the gradient and expectation

$$\nabla_\phi \text{ELBO}(\phi) = \nabla_\phi \mathbb{E}_{q(\epsilon)}\left[\log \frac{p(\mathcal{T}_\phi(\epsilon), x)}{q_\phi(\mathcal{T}_\phi(\epsilon))}\right] = \mathbb{E}_{q(\epsilon)}\left[\nabla_\phi \log \frac{p(\mathcal{T}_\phi(\epsilon), x)}{q_\phi(\mathcal{T}_\phi(\epsilon))}\right]$$

$$\nabla_\phi \int q(\epsilon) \log \frac{p(\mathcal{T}_\phi(\epsilon), x)}{q_\phi(\mathcal{T}_\phi(\epsilon))} d\epsilon = \int q(\epsilon) \nabla_\phi \log \frac{p(\mathcal{T}_\phi(\epsilon), x)}{q_\phi(\mathcal{T}_\phi(\epsilon))} d\epsilon$$

Reparameterization Gradient Estimate

This gives a simple unbiased Monte Carlo estimate of the gradient:

$$g = \nabla_\phi \left(\frac{1}{K} \sum_{i=1}^{K} \log \frac{p\left(\mathcal{T}_\phi(\epsilon^{(i)}), x\right)}{q_\phi\left(\mathcal{T}_\phi(\epsilon^{(i)})\right)} \right), \qquad \epsilon^{(1)}, \ldots, \epsilon^{(K)} \sim q(\epsilon)$$

We can compute it as follows:

1. Draw $\epsilon^{(1)}, \ldots, \epsilon^{(K)} \sim q(\epsilon)$
2. Compute $\widehat{\text{ELBO}}(\phi, \epsilon^{(1)}, \ldots, \epsilon^{(K)})$ = term in parentheses above
3. Use autodiff to get $g = \nabla_\phi \widehat{\text{ELBO}}(\phi, \epsilon^{(1:K)})$

backpropagation, JAX, PyTorch

Reparameterization with Gaussians

$\mathcal{N}(\mu, \Sigma)$

$d \times d$

Multivariate Gaussians are common variational distributions, and easy to reparameterize:

$$\epsilon \sim \mathcal{N}(0, I) \Longrightarrow \mu + L\epsilon \sim \mathcal{N}(\mu, LL^\top).$$

This is a reparameterization with

$$q(\epsilon) = \mathcal{N}(\epsilon|0, I), \qquad \mathcal{T}_\phi(\epsilon) = \mu + L\epsilon \qquad \phi = (L, \mu)$$

It covers any multivariate Gaussian, since an arbitrary covariance matrix Σ can be written as $\Sigma = LL^\top$ for some L (e.g., a Cholesky factor)

Reparameterization with Diagonal Gaussians

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{bmatrix}$$

Another common variational distribution is a diagonal Gaussians:

$$q_\phi(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)) \qquad \sigma^2 = [\sigma_1^2, \ldots, \sigma_d^2]$$

This can be reparameterized as:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

$$z_j = \mu_j + \sigma_j \epsilon_j$$

($\odot$ = elementwise multiplication)

Example: Factor Analysis

Let's return to our Bernoulli VAE model

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, I)$$
$$p(x_j|\mathbf{z}) = \text{Bernoulli}(x_j; \underbrace{(f_\theta(\mathbf{z}))_j}_{\mu_j}), \qquad j = 1, \ldots, d$$

Suppose we choose a diagonal Gaussian variational family

$$q_\phi(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mu, \text{diag}(\boldsymbol{\sigma}^2))$$

We now know how to use BBSVI with the reparameterization trick to optimize the ELBO.

With the optimized parameters $\phi = (\mu, \boldsymbol{\sigma})$ we can

- approximate $p(\mathbf{z}|\mathbf{x}) \approx q_\phi(\mathbf{z})$
- lower bound the log-marginal likelihood $\log p(\mathbf{x}) \geq \text{ELBO}(\phi)$

Next time: learning the model f_θ. Thoughts?

HW5

$f_\theta(z) = Wz$, "model" $\theta = W$

Bonus: Handling Some Terms in Closed Form

The ELBO can be decomposed into several terms with different computational properties:

$$\begin{aligned} \text{ELBO}(\phi) &= \mathbb{E}_{q_\phi}\left[\log \frac{p(Z,x)}{q_\phi(Z)}\right] \\ &= \underbrace{\mathbb{E}_{q_\phi}[\log p(Z)]}_{\text{"cross entropy"}} + \underbrace{\mathbb{E}_{q_\phi}[\log p(x|Z)]}_{\text{"energy"}} - \underbrace{\mathbb{E}_{q_\phi}[\log q_\phi(Z)]}_{\text{"entropy"}} \end{aligned}$$

With simple distributions (esp. Gaussians) the cross entropy and entropy terms can often be computed in closed form.

Example: cross entropy standard normal and diagonal Gaussian

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}; 0, I) \\ q_\phi(\mathbf{z}) &= \mathcal{N}(\mathbf{z}; \mu, \mathsf{diag}(\boldsymbol{\sigma}^2)) \end{aligned} \implies \int q_\phi(\mathbf{z}) \log p(\mathbf{z}) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{d}(\mu_j^2 + \sigma_j^2)$$

When possible, it's usually (but not always) best to compute these terms and their gradients analytically, and only use Monte Carlo estimation for the energy term.

This is because lower variance gradient estimates will make the stochastic optimization converge faster.